文本分类技术在报业智能客服系统中的应用

郑创伟 谢志成 邢谷涛 陈少彬 陈义飞
（深圳市创意智慧港科技有限责任公司，广东 深圳 518034）

摘 要：作为主流媒体的新闻报刊业，从广告办理、报纸印刷、发行订报到新闻报料，每天接收和处理的客户信息量都巨大。近年来，随着人力成本的日益增长，传媒行业对可以大幅减轻人工客服工作量的智能客服系统的需求也日益迫切。本文从提高报业行业人工智能客服系统的准确率和效率的目的出发，对问句分类技术进行了深入研究，主要使用快速文本分类算法（fastText）来实现对问句的分类。实验表明，与传统的 SVM+TF、BERT+SVM 相比，该算法能够很好的兼顾系统的查询准确度和查询的时间开销以满足用户的查询准确率和速率的需求。

关键词：智能客服系统；问句分类技术；fastText；分层 softmax　　**中图分类号：**G211　　**文献标识码：**A
文章编号：1671-0134（2021）10-149-03　　DOI：10.19483/j.cnki.11-4653/n.2021.10.045
本文著录格式：郑创伟，谢志成，邢谷涛，陈少彬，陈义飞．文本分类技术在报业智能客服系统中的应用 [J]. 中国传媒科技，2021（10）：149-151.

1. 智能客服系统的概念

智能客服系统是一个涉及多种先进技术的综合体，例如：大规模知识处理技术、自然语言理解技术、知识管理技术、自动问答系统、推理技术等，不仅为企业提供了细粒度知识管理技术，还为企业与海量用户之间的沟通建立了一种基于自然语言的快捷有效的技术手段；同时还能够为用户提供精细化管理所需的统计分析信息。

2. 发展背景

随着多媒体技术的不断成熟，多媒体信息接入渠道从传统的网点、电话、网站、邮件到即时通信、微博、微信等不断涌现，网络信息呈现出碎片化、移动化、实时化、个性化、多媒体化、大数据化的特点。这些变化给信息管理和服务带来全新的挑战，报业传统的客服系统已经满足不了大众的服务需求。

与此同时，人工智能领域的智能机器人技术发展迅速，对生产生活方式产生革命性的影响。针对报业领域，考虑到目前互联网信息爆炸式增长现状，使用智能机器人技术可以大幅减少人工客服的工作量，有效提高信息处理效率。因此，智能机器人应用到报业客服服务中已成为必然趋势。

3. 相关工作

近几年来，文本分类技术被广泛的应用在智能客服系统中。传统的分类算法有多层感知器、朴素贝叶斯和支持向量机（SVM）等，其中 SVM 是一类按监督学习方式对数据进行二元分类的广义线性分类器。其优点是在优化问题的同时考虑了经验的风险和结构风险最小化，因此具有一定的稳健性，而铰链损失函数的取值特点使 SVM 具有稀疏性。缺点是超过一定量的数据时，训练的时间呈指数增长。因此，当数据集的量过大时，传统的分类算法就不具有优势，反正成为了降低其运算效率的最主要原因。[1]

大数据时代的到来使得深度学习的分类算法快速发展，例如 TextCNN 是由 Yoon Kim 在 2014 提出的，它的结构和图像处理的过程非常相似，是由一个卷积层和一个最大池化层组合的结构，输入是一个词向量矩阵，卷积层使用不同宽度的卷积核在整个句子长度上滑动，得到 n 个激活值，然后再通过最大池化层得到 m 个特征值组成的 feature map 来供后级分类器作为分类的依据。[2]

在速度要求特别高的场景中，fastText 的算法是一个不错的选择，它不仅能保证以较高的速度完成模型训练，还能保证模型的精度。[3] 本文研究使用 fastText 作为智能客服系统中的领域分类器，并与传统的分类算法做对比实验。

4. 快速文本分类算法（fastText）

该种分类算法的一个简单而有效的基准是将句子表示为单词袋，并且可将一篇媒体报道中的单词送入一个线性分类器中，并训练该线性分类器。然而，线性分类器不能在特征和类之间共享参数，这可能限制了泛化。这个问题的常见解决办法是将线性分类器分解为低秩矩阵或使用多层神经网络。使用神经网络，可以通过隐藏层来共享信息。fastText 方法包含三部分，即模型架构、层次 softmax 和 N-gram 特征。fastText 方法将单词序列作为输入，通过利用 softmax 函数来对单词类别的概率分布情况进行预测，这是一种使用随机梯度下降和反向传播进行模型训练的方法。

4.1 模型架构

fastText 的模型架构和 word2vec[4] 中的 CBOW 模型的结构相似。CBOW 模型是利用上下文来预测中间词，而 fastText 是利用上下文来预测文本的类别。而且从本质上来说，word2vec 是属于无监督学习，fastText 是有监督学习。但两者都是三层的网络（输入层、单层隐藏层、输出层），具体的模型结构如下。

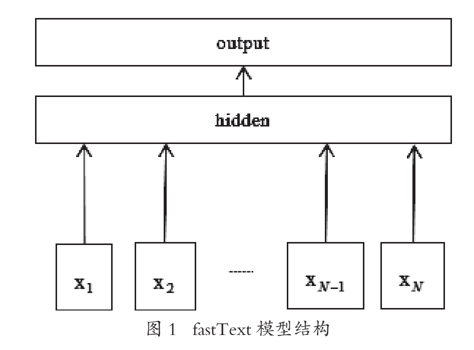


图 1 fastText 模型结构

图中 x_i 表示的是文本中第 i 个词的特征向量。该模型将一系列单词作为输入并产生一个预定义类的概率分布，且在多个 CPU 上以异步方式进行训练，所以训练的速度非常快。然后通过一个 softmax 方程来计算这些概率，以霍夫曼编码树来建立方程，树中的每个节点都与从根节点到该节点的路径概率相关联，公式如下。

$$p(n_l+1)=\prod_{i=1}^{l}p(n_i)$$

可以看出，某个节点的概率是小于其父节点概率的，对整个霍夫曼编码树进行深度优先搜索并找到节点之间的最大概率，这样就能够通过舍弃一些小概率分支来降低复杂度。

4.2 分层 softmax

当数据的类别很多时，线性分类器的计算量将会大增。其计算复杂度为 $O(Nd)$，其中 N 是分类的数量，d 是隐藏层的维度。为了降低计算量，该模型使用了一个基于霍夫曼编码树的分层 softmax（而非扁平式架构），它属于逻辑回归在处理多类别任务上的推广。不同的类别被整合进树形结构中，同时也考虑到类别不均衡（一些类别出现次数更多）的现象。[5] 根据图 2 示意，叶子节点由底部 K 个不同的类标组成，K–1 个内部节点作为内部参数，其公式表示如下：

$$\mathrm{P}(y_j)=\prod_{l=1}^{L(y_j)-1}\sigma\left(\left[\!\left[n(y_j,l+1)=LC\left(n(y_j,l)\right)\right]\!\right]\cdot\theta_{n(y_j,l)}^{T}X\right)$$

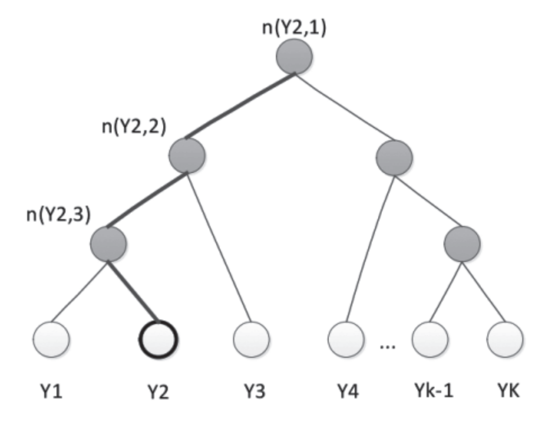


图 2 分层 softmax 示例图

通过使用霍夫曼编码树建立用于表征类别的树形结构，因此，频繁出现类别的树形结构的深度要比不频繁出现类别的树形结构的深度要小，这也进一步的提高了计算效率。在训练过程中，计算的复杂度会下降到 $O(d\log_2(N))$ 。

4.3 N–gram 特征

不管是文本分类还是句子分类，常用的特征是词袋模型。但词袋模型不能考虑词之间的顺序，因此该模型还加入了 N–gram 特征，其基本思想就是将客户咨询的文本内容按照字节进行大小为 N 的滑动窗口操作，从而得到多个长度为 N 的字节片段序列。并且为了提高效率，还需过滤掉低频的 N–gram，按照事先设定好的阈值形成关键列表。[6] 通过 N–gram 获取局部词序信息，其用于评估顾客咨询的语句是否合理，从而能够更加平衡获取词序信息及计算资源。

在实际报业智能客服应用中，客户咨询产生的大量文本数据，在对其进行分类时，N–gram 处理后生成的词条会存在大量冗余，因此必须对处理后的词语进行重构，区分出客户实际想要咨询的问题，在此过程中，也能够学习到更多词语与词语之间的前后关联特征。在对词语进行重构的过程中，不考虑客户所表述词语的语法和词序之间的关系，也就是每个词都是相对独立的，然后对比词袋中的词条和 N–gram 处理后的结果，最后将文本词袋中没有而 N–gram 处理后存在的词条删除。

5. 实验结果及分析

在本节中，将使用 fastText 与传统的 SVM+TF 和 SVM+BERT（使用 BER[6] 提取词向量）做对比实验。主要的目的是体现 fastText 在数据量足够多时以及分类的类别比较多的情况下的优势。

5.1 二分类

采用智能客服系统线上数据做训练和测试，数据的类别有两类，即闲聊和服务咨询。首先对线上数据进行预处理，预处理的目的主要是去除不符合条件的数据。预处理后，从数据库中得到闲聊和服务咨询各 5 万条，实验时，从中选取 1 万条测试数据。本次测试以测试数据的准确率、模型训练的时间以及 ROC 图（ROC 和 AUC 是评价二分类器的指标）作为二分类模型评判的标准。实验结果如表 1 所示：

表 1 算法实验结果对比 1

Method	Accuracy	Train time（s）
SVM+TF	99.81%	9.33
SVM+BERT	99.47%	1677.89
fastText	99.60%	13.2 1

通过表 1 的实验结果可以发现 SVM+TF 算法准确率要略高于其他两种算法，模型训练的时间要低于其他两种算法。其中 SVM+BERT 算法训练时间太长，不适用于

当前应用场景，所以下述实验不再对比 SVM+BERT 算法，只考虑 SVM+TF 和 fastText 算法，对此两者继续实验。

紧接着上述实验，增加实验的数据集，对比 SVM+TF 和 fastText。数据量由原来的 10 万增加到 165 万，并按 9：1 的比例分割训练集和测试集。实验结果如表 2 所示。

表 2 算法实验结果对比 2

Method	Accuracy	Train time（s）	F1–score	
			Chat	Service
SVM+TF	99.74%	162.02	98.08%	99.93%
fastText	99.62%	121.3　8	99.46%	99.76%

从表 2 可以得知总体的准确率都有所提升，而 fastText 算法准确率要略低于 SVM+TF，但其训练模型的时间要远远低于传统的 SVM+TF 算法。从训练时间的角度来看，fastText 具有一定的优势。下图是 SVM+TF 和 fastText 算法的 ROC 图。

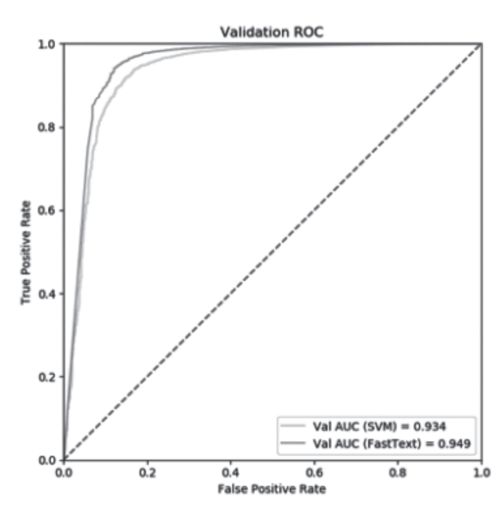


图 3　SVM+TF 和 fastText 算法的 ROC 图

从上图中可以得出 fastText 的 AUC 的值要大于 SVM 的。综合表 2 和图 3 可以得知 fastText 在大数据量的条件下，分类效果和 SVM 基本一致。

5.2　多分类（三分类）

SVM 是一个线性分类器，对于二分类问题，使用 SVM 是特别方便的，但对多分类，SVM 就不如 fastText 方便。本实验中，使用 SVM 实现三分类采用的是间接法，即使用两个二分类器。将上述数据分为三个类别即闲聊、地理服务和咨询。实验结果如表 3 所示。

表 3　算法实验结果对比 3

Method	Accuracy	Train time（s）	F1–score		
			Chat	Route	Qa
SVM+TF	99.10%	231.85	98.04%	99.88%	99.97%
fastText	99.71%	115.53	99.69%	99.38%	99.72%

从上表可以看出，无论是从准确率和模型训练时间来看，fastText 算法都要优于 SVM。

6. 总结

综上所述，在智能客服实际应用中，当客户咨询量较少或数据量较少时，问题复杂度低（类别少）的情况下，SVM 做分类的效果要比 fastText 算法效果好。但当客户咨询量增大或分类类别变多时，fastText 做分类就开始显示出一定的优势，其不但分类效果好，而且速度非常快，非常适合于智能客服的应用。这也就意味着，随着当前用户规模的不断上升，报业在实际运营过程中将面临更多客户咨询问题，因此所产生的数据量也呈现爆发式增长，因此在未来报业智能客服系统实际应用中，可以选择 fastText 算法进行建模和尝试。媒

参考文献

[1] 谢季川，宗振国，刘宏国，张春秋，田晓 . 基于词向量模型的 95598 工单文本挖掘 [J]. 电子世界，2017（23）：176+178..

[2] 冯梦莹，李红 . 文本卷积神经网络模型在短文本多分类中的应用 [J]. 金融科技时代，2020（1）：38–42.

[3]Joulin A，Grave E，Bojanowski P，et al. Bag of Tricks for Efficient Text Classification[J]. 2016.

[4] 熊富林，邓怡豪，唐晓晟 .Word2vec 的核心架构及其应用 [J]. 南京师范大学学报（工程技术版），2015（1）：43–48.

[5] 李颢，张吉皓 . 基于文本挖掘技术的客服投诉工单自动分类探讨 [J]. 移动通信，2017（23）：66–72.

[6]Devlin J，Chang M W，Lee K，et al. BERT：Pre–training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.

作者简介：郑创伟（1978–），男，广东汕头，高级工程师，研究方向：大数据、人工智能；谢志成（1980–），男，广东汕头，中级职称，研究方向：大数据、云计算；邢谷涛（1984–），男，海南文昌，中级，研究方向：云计算；陈少彬（1973–），男，广东揭阳，中级，研究方向：大数据；陈义飞（1981–），男，广东湛江，中级，研究方向：大数据。

（责任编辑：陈旭管）